A Class-Based Language Model[1]

Main Idea: cluster words into a fixed number of clusters C and use their cluster assignments as their identity instead (reducing sparsity)

If $\pi : V \to C$ is a mapping function from a word type to a cluster (or class), we want to find

$$\pi^* = \arg\max p(\mathbf{W} \mid \pi)$$

where

$$p(\mathbf{W} \mid \pi) = \prod_{i=1}^{N} p(c_i \mid c_{i-1}) p(w_i \mid c_i)$$

with $c_i = \pi(w_i)$.

[1] Peter F. Brown et al. "Class-based N-gram Models of Natural Language". In: *Comput. Linguist.* 18.4 (Dec. 1992), pp. 467–479.